Chip-to-Cloud: an Autonomous and Energy Efficient Platform for Smart Vision Applications

A. Scionti, S. Ciccia, O. Terzo, G. Giordanengo
Istituto Superiore Mario Boella (ISMB)
Torino, Italy
{scionti,ciccia,terzo,giordanengo}@ismb.it

***Abstract*—Modern Cloud architectures encompass computing and communication elements that span from traditional data center computing nodes (offering almost infinite resources to satisfy any application demands) to edge-computing and IoT devices (to sense and act on the real world). This paper presents the Cloud architecture devised within the OPERA project [1], which provides new levels of energy efficiency as a full chip-to-Cloud solution. Focusing on a smart vision application (i.e., road traffic monitoring), the paper presents novel architectural solutions optimised to achieve high energy efficiency at any level: *i*) the computing elements supporting the acceleration of State-of-the-Art CNNs; and *ii*) an innovative wireless communication subsystem. Unlike conventional designs, our wireless communication subsystem exploits the advantages of software defined radio (SDN) firmware to control a reconfigurable antenna. To further extend the application range, an energy harvesting module is used to supply power. Besides the edge-IoT, high-density accelerated servers offer capabilities of running complex algorithms within a small power envelop. The effectiveness of the whole architecture has been tested in a real context (i.e., 2 installation sites). In-field measurements demonstrate our claim: high-performance coupled with high energy efficiency over the whole system.**

***Index Terms*—cyber-physical systems, hardware acceleration, wireless communication**

I. Introduction

Continuous progresses in silicon technologies and processing architectures represent the key elements enabling the creation of new smart systems. Improved processing capabilities open the door to devices (Cyber-Physical Systems –CPS) that sense, make decisions by running complex algorithms, and act on the physical environment. The large availability of such devices at low cost, and their low power consumption, enables end users to deploy them on large scale. Modern Cloud architectures include CPS at the edge, to enable more flexible distribution of computing capabilities, as well as to close the gap with applications requiring near-real time processing. To this purpose, mechanisms to ease the interaction between the edge nodes and the data center core are used [2], [3]. For instance, public Cloud platforms include specific API and dedicated frameworks for programming and managing edge devices, as well as tasks' offloading.

Moving from CPS to data center (DC), processing capabilities and power budget change: CPS make large use of optimised architectures (system-on-chips – SoCs), which provide the right level of performance in a (ultra-)low power envelop, also enabling energy harvesting solutions to be used. Indeed, specialised IPs are used to accelerate critical functionalities of the system. While large effort in past research works has been spent in designing and optimising processing elements, less effort has been devoted to the communication aspects. However, large deployments of CPS require the creation of ad-hoc networks (e.g., wireless sensor networks –WSNs). Thus, communication, especially if based on wireless technologies, plays a key role in determining the overall energy efficiency of the system. CPS need to exchange data with other nodes and DC core, thus a not-optimised radio communication subsystem may quickly consume the largest part of the power budget.

Video surveillance represents a well established application domain addressed both at the edge and core of Cloud architectures. When addressing traffic monitoring applications, energy efficiency of camera sensors at any level becomes mandatory. While these are demanding for high data rates (e.g., video dispelling doubt) which further increase the radio energy need, reducing the energy consumption is crucial since camera sensors are mostly supplied by renewable sources and small batteries [4]. Similarly, to ensure energy efficiency in any point of the architecture, an energy efficient engine is also required in the DC core. For instance, whenever the camera sensors detect specific events, the video streaming is passed to the DC core where more powerful algorithms are used to analyse the events.

This paper describes the solution devised and implemented within the EU funded OPERA project [1]. By relying on high efficiency processing technologies (i.e., ultra-low power SoCs and FPGAs), as well as on reconfigurable antennas for the communication subsystem, the presented solution provides high flexibility and performance, still guaranteeing energy autonomy (i.e., a harvesting solution based on a solar panel). Flexibility and adaptability is obtained through a complete software approach: the firmware controlling the antenna has been optimised to take full advantage of the capabilities of the embedded system without impacting on the application performance (i.e., the camera sensor still performs on-board image processing and video streaming towards the DC core).

II. Traffic Monitoring Platform

The general architecture designed for the traffic monitoring case study is depicted in Figure 1. Being devised to support image processing both at the edge and in the DC core,

978-3-9819263-2-3/DATE19/©2019 EDAA

the architecture envisages three main elements: *i*) an ultra-low power computing device (ULP) equipped with a camera sensor, *ii*) an advanced radio communication subsystem (i.e., the software radio controller and the directive antenna) which is directly connected to the ULP, and *iii*) a remote low power server equipped with an FPGA board for accelerating complex image processing tasks. In addition, an end-point is placed between the ULP and the remote server. This latter has the purpose of receiving data from ULPs and forwarding them to the remote server through a standard wired link. It is worth noting that wireless communication is essential to support communication in those area not reached by wired infrastructures. Also, it helps to cover larger geographical areas through multiple nodes (e.g., WSNs).

The ULP is equipped with an ultra energy efficient computing solution. To this end, the system embeds an efficient general purpose micro-controller along with several DSPs and dedicated hardware function accelerators. Specifically, it sports hardware blocks for accelerating the execution of convolutional neural networks (CNNs), which in turns are used for processing acquired images. Similarly, the remote server is provided with a dedicated FPGA accelerator, that implements custom processing pipelines for more sophisticated CNNs. This latter has been efficiently synthesised by relying on a HLS compilation tool-chain, which accelerated its development and test.

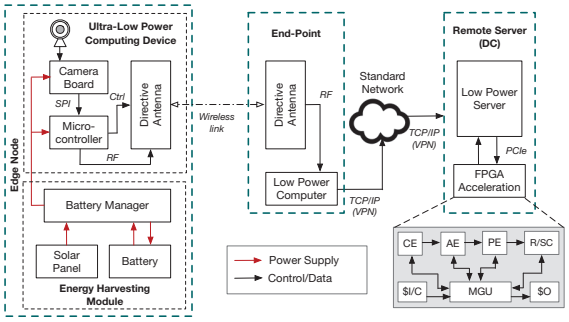

Fig. 1. Overview of the proposed chip-to-Cloud architecture for traffic monitoring applications: black arrows represents control/data signals, while red arrows shows power supply paths.

The capability of optimising the radio communication subsystem strictly depends on the communication protocol to support. WiFi, is a candidate to provide high transfer rates with low power consumption. However, standard implementations of such protocols offer a covered distance that is in general not enough for real applications, such as road traffic monitoring. To overcome this limitation, modulation schemes consisting of multiple transmitters and multiple receivers are used; however, the power consumption of such solutions make them unfeasible for edge applications.

According to this premise, the 802.11b/g standard was selected as the best candidate for our purposes; although, radio communication link must be optimised in order to greatly extend the covered distance. To this end, a software controlled directive antenna has been implemented. In principle, a system equipped with reconfigurable directive antenna has the following advantages in term of energy efficiency with respect to a standard (omnidirectional) one. First, given a certain received signal power, the transmission power is reduced. Thus, without changing the communication performance the system transfers data with less energy. Conversely, given a certain power of transmitted signal, the received signal power is higher. As a consequence, the system can transmit data at faster rates at a longer distance. Again, less energy is required since the information can be transferred in a shorter time.

Looking the configuration of the edge node, the following key components are found: *i*) a video sensor board which embeds the software for video detection; *ii*) a secondary micro-controller which acts as an interface between the video sensor and the radio subsystem; *iii*) the radio IEEE 802.11b/g module, which provides wireless connectivity. Finally, a reconfigurable antenna which could be of the type of a phased array (fine scan within 90° of coverage) or a beam-switchable antenna (360° switching capability) is present. The secondary micro-controller holds the firmware controlling the antenna radiation. The low energy consumption of the entire edge node allows the exploitation of an energy harvesting subsystem, which supplies the autonomous wireless sensor. The harvesting subsystem consists of: *i*) a photovoltaic module; *ii*) a battery manager module that controls and optimises stored energy levels; and *iii*) a battery module which stores excess of energy (when radiation produces more energy than needed) and to feed energy when radiation is not enough. Finally, an end-point (gateway –GW) is used to transfer data received by ULP nodes to the remote server. It is composed of a radio communication subsystem (i.e., the receiver, which is identical to that used by ULP transmitter), and a single board computer embedding a SDN receiver to collect on-air data and flushing these data over Internet [5]. The interaction between the edge nodes and the Cloud server is based on an ad-hoc framework. On the ULP node side, the application embeds and runs a dedicated software component to offload image processing when such tasks demand for large computing resources; on the Cloud server side, a service Linux container is in charge of receiving remote processing requests and offloading them on the FPGA acceleration pipeline.

III. Reconfigurable Antennas

Of particular interest for improving energy efficiency of the radio communication is the antenna configuration process, i.e., the ability of directing the beam toward the desired direction in order to guarantee the advantages discussed in Section II. The most important factor influencing the antenna design is represented by the link distance to cover. On long distances, *phased arrays* would be able to provide a finer antenna alignment. However, the scan range is limited. On the other hand, *switchable antennas* would provide an overall scan range of 360°, but with lower precision. This paper will address this compromise, discussing two validation cases.

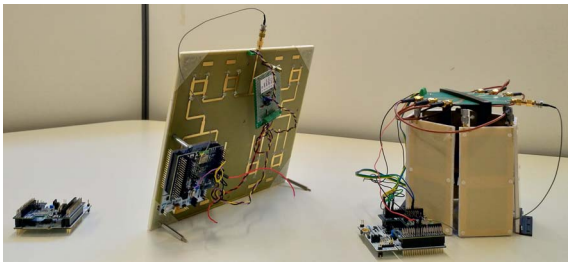

Fig. 2. From left to right, State-of-the-art wireless module (built-in antenna), steerable antenna prototype, switchable antenna prototype

A. Long-range communication: finer scan

In a first site, the camera system has been installed on a top of a streetlight pole, far 1 Km from the gateway. In this conditions, an autonomous antenna configuration mechanism (i.e., electronic beam steering) is required to enhance the quality of the link and reach the required communication distance.

To overcome existing limitations, the developed antenna is composed by two parts: an array of four patches and the Beam Forming Network (BFN). This latter is composed by analog phase shifters which allow to steer the main beam in the desired direction. Since phase shifter are controlled by analog voltages, the main beam can assume any direction in a scan sector of $90°$ providing a very fine scan. Any direction pointed by the beam requires some processing to analyse the channel. The firmware running on the micro-controller works in the following manner. For each searched direction, the beam is pointed; then, the radio transmits a beacon frame and acquires the result to make a decision. Receive Signal Strength Indication (RSSI) and transmission rate information are available in the beacon frames. Algorithm 1 shows the main operations of the fine scan process. Although the analysis

Algorithm 1 Antenna controller for steerable antenna.

```
for any direction from −45° to +45° do
    set(analog values to configure the beam direction)
    send(beacon frame)
    receive(beacon feedback)
    store(RSSI value for this direction)
end for
search(best RSSI direction)
set(the beam in the best direction)
```

of each link direction requires less than 1 second, to avoid unnecessary latency, the number of directions to look for has been reduced. Since the proposed antenna has a beam that covers a sector of $30°$, an appreciable change of 3.0 dBm in RSSI is sensed by steering the main beam with steps of $15°$. As a result, only 7 directions (i.e., $-45°$, $-30°$, $-15°$, $0°$,$15°$, $30°$, $45°$) are necessary to analyse the whole scan plane. To favour speed on field applications, we further limited the scan to only 5 directions (see Table I). An interesting point is that, the state-of-the-art radio systems were not able to communicate at such link distance (i.e., 1 Km), even at maximum transmit power. Reconfigurable antenna demonstrates a reliable communication with an improved RSSI (-78 dBm), and a transmit power configured at half dynamic. Thus, also the power consumption of the radio is reduced when transmitting.

B. Extending the configuration process to $360°$

In a second installation site we addressed the problem concerning the limited beam steering of phased array antenna. Specifically, with a standard steerable beam antenna, the wireless node can direct its beam toward the desired direction with a fine precision, albeit with a limited scan range. The improved version of the antenna removes this limitation by showing a $360°$ beam switching capability.

The designed antenna is composed of six sub-arrays organised in a 3D fashion and forming an hexagon (see Figure 2). Each sub-array points in a different direction and they are activated one at a time by a RF-switch. The switch allows to form six radiation patterns in different azimuth directions (i.e., $0°$, $45°$, $135°$, $180°$, $225°$, $270°$). Compared to the phased array antenna, in this case, the beam is $65°$ large, thus providing a coarser scanning capability. In-field test and average RSSI measured for the $360°$ scan system is reported in Tab. I (multiple measurements have been performed, and the average value is presented).

TABLE I
RSSI [*dBm*] FOR EACH POINTING DIRECTION OF THE ANTENNA BEAM.

Steerable antenna						
Direction $\theta[°]$	$-40°$	$-20°$	$-0°$	$20°$	$40°$	-
Avg. RSSI	-93	-92	-92	-79	-77	-
Switchable antenna						
Direction $\theta[°]$	$0°$	$45°$	$135°$	$180°$	$225°$	$270°$
Avg. RSSI	-46	-40	-52	-64	-71	-62

IV. Image Processing Acceleration

The need of processing acquired images in real time led to the implementation of an advanced CNN running on the SoC embedded in the ULP device. Compared to standard PCIe-based acceleration board, the SoC has limited resources; although several dedicated hardware function accelerators are present (e.g., colour converters, encoders, image cropper, etc.). Furthermore, a set of dedicated accelerator for the main CNN operations is available (e.g., CNN kernel compression, reconfigurable data transfer fabric to improve data reuse,etc.). The software running on the ULP has been split in two parts: *i*) the CNN code devoted to detect objects in the scene, and *ii*) the code that uses the results of the CNN detection to implement higher level functionalities such as filtering, tracking and object counting (e.g., counting bicycles in a group). The implemented CNN is customised version of state-of-the-art CNN architecture, which has been optimised to fit with the limited resources available on the chip. To this end, the CNN architecture has been derived from YOLO [6], by reducing the number of layers to fit in the chip memory. As

such, embedded hardware accelerators and the DSP cluster have been fully exploited. The CNN has been trained with a subset of the COCO data set (i.e., bicycles, motorbikes, car, trucks, bus, pedestrians have been considered as classes). Whenever a new frame (i.e., an image) is captured by the camera sensor, a set of pre-filtering operations are applied. After filtering and object detection process, the system software uses the results of detection to identify any object previously detected (tracking). Tracking is essential in order to correctly counting objects in the scene that moves in as a group (e.g., a group of bicycles), and to correctly detect any change in their number. On the other hand, complex tasks such as object detection and classification of complex scenes requires algorithms demanding more computing resources than those available on the single ULP device. In the road traffic monitoring context it is important to detect critical situations as soon as possible. For instance, knowing the number of bicycles in a group moving on the road is important to quickly detect accidents and raise alarms. Discrete accelerators running on data center servers offer the capability of achieving high accuracy levels of detection and classification, as required to track and count multiple objects in a group. To this purpose, we devised a solution implementing (on a mid-range reconfigurable device –Intel Arria10) a dedicated acceleration pipeline for a state-of-the-art CNN (YOLO) [6], tailored for on-line image processing. The pipeline consists of the following computing engines (conveniently associated to running FPGA-kernels): *i*) convolution engine (CE); *ii*) activation function engine (AE); and pooling/normalisation engine (PE); a routing (R) and a shortcut (SC) blocks (which are used by residual operations. Two additional stages are present to load input image channels and network weights. To reduce the pressure on the memory subsystem, internal caches have been implemented using distributed memory blocks. The synchronisation among the stages of the pipeline is provided by a dedicated management unit (MGU). This pipeline has been described and integrated into the CNN/application code, by exploiting high-level synthesis tools (HLS Intel OpenCL compiler). To maximise performance, input data is sliced and cached locally in FPGA memory blocks, and coefficients are double buffered and loaded during calculation. The output is also cached internally.

V. Energy Harvester

An energy harvesting module is very important for a wireless monitoring system since the energy grid is not always available, especially in rural areas. Photovoltaic technology was preferred with respect to wind turbines as well as other technologies (e.g., thermoelectric, vibrational and wireless power transfer) since, simulation results [7] demonstrated that for the selected test sites, solar radiation is the most available source. Designing the harvester module requires to define the overall power consumption required by the target system (i.e., the camera board and the radio) to perform its operations:

$$P_L = \frac{\sum_i P_{case}(i)}{24h} + P_{camera}\ [W] \quad (1)$$

where P_{camera} is the camera board power consumption, and $P_{case}(i)$ accounts for the power consumed by the radio module to transmit data related to each test case in one day (it is worth to note that the harvester module is the same for all the test sites). The latter is described as follows:

$$P_{case}(i) = t^i_{TX} \cdot P^i_{TX} + (24h - t^i_{TX}) \cdot P_{idle}\ [Wh/day] \quad (2)$$

where P_{TX} is the power consumption of the radio module when transmitting, P_{idle} is the power consumption when the radio is not used, and $t_{TX} = (\frac{S}{R}) \cdot 0.28x10^{-3}\ [h]$ is the overall time required to transmit all the information in one day. Here, R is the wireless data rate employed by the radio module expressed in bit per second (bps), while S is the amount of data to transmit daily. Such amount depends on the size of the alarm event (S_a), and the time (t_v) required to send the related video stream, which in turn relates to the image size (S_i) of each transmitted frame, and the frame rate F_R. All these factors are combined with the number N_e of events statistically foreseen, as follows $S = N_e \cdot (S_i + t_v \cdot (S_a \cdot F_R))\ [b]$.

Based on the load power consumption (P_L), sizing the solar panel can be done through the following equation:

$$P_{PVmin} = \eta\ 24h\ P_L(h_{eq})^{-1}\ [W_p] \quad (3)$$

where $\eta = 1.84$ is a correction factor used as a safety margin in the calculation to compensate the power value due to the variable weather conditions and other loss factors (e.g., dust, ice, snow, etc.). Finally, h_{eq} is the equivalent hour parameter for a typical day of poor radiation and is expressed as the ratio between H_{60°, i.e., the irradiation on plane at 60° in the worst month of the year [7], and P_{STC}, i.e., the light intensity (which is considered $1000W/m^2$ on the whole surface of the module). This latter is about what is available at noon on a sunny day when the module is facing towards the sun [8].

The minimum capacity of the battery, without considering losses, can be computed as follows:

$$C_{min} = h_{aut} P_L(V_b)^{-1}\ [Ah] \quad (4)$$

where h_{aut} is the battery duration, and V_b is the required voltage to supply the whole ULP system. Since the discharge cut-off limit has to be accounted when sizing the battery, a corrective factor of 40% is added to C_{min}.

The requirements and statistics of expected number of detection and video duration are summarised in Table II. As a rule of thumb, the harvester sizing is estimated in the worst case scenario. Therefore, we consider the radio module configured at maximum transmission power (i.e., $P_{TX} = 1.14$ W) and minimum wireless data rate (i.e., $R = 1$ Mbps), while we set $P_{idle} = 0$ W. The size of an alarm is considered to be 320 Kb, while for the spelling doubt, the frame rate is $F_R = 8$ fps. From equation (1), the total power consumption for the radio module equals to 0.08 W. Conversely, the camera board is always active to continuously processing data, and its average power consumption has measured as $P_{camera} = 0.5$ W. Thus, the power consumption of the

load equals to $P_L = 0.58$ W (Equation (1)). The size of the solar panel can be estimated via Eq. (3) by setting $H_{60^\circ} = 1880$ Wh/m^2/day (Grenoble –France, the location of installations). This gives a $P_{PVmin} = 14$ Wp. We selected a commercial photovoltaic module characterized by 20 Wp. The solar panel has a dimension of 60 cm$\times$20 cm [9]. The battery to store the energy has a requirement of 1 week with a nominal voltage of 12 V. Given these constraints the minimum battery capacity equals to $C_{min} = 7.2$ Ah (accounting losses). We opted for a commercial battery of 10 Ah [10].

VI. Experimental Validation

In the following the efficiency of the proposed solution is discussed. We measured the efficiency of the whole chip-to-Cloud on the field (two installation sites have been set up for this purpose, in Grenoble – France), by assessing the power(energy) saving allowed by the improved wireless communication subsystem, and the efficiency of the image processing solutions.

An interesting point is that on the first validation site, thanks to the antenna configuration process we were able to reduce the signal transmission power at half of its dynamic. This reduced the power consumption when transmitting by 20%. On the other hand, the energy harvester module was designed to guarantee up to 7 days of full operations for the ULP device. Again, thanks to the reduced power consumption, also the autonomy of the system is extended. Similarly, installation on the second site showed large power saving on data transmission. Effectiveness of the wireless transmission system have been also demonstrated by the sustained bit-rate, which was very close to the theoretical one (depending on the implemented standard; e.g., close to $\sim$ 54 Mbps with IEEE 802.11 standard). Similarly, transmission distance reached up to 1 km on in-field tests (first installation site).

The ULP device is based on an advanced STMicroelectronics platform [11]; the CNN implemented on the ULP side exploited the convolutional hardware accelerators as well as the embedded DSPs. The platform was designed to achieve best in-class energy efficiency and performance trade-off, offering up to 2.9 TOPS/W.

For the specific application, the performance (of the image processing pipeline implemented on the remote FPGA – see Section IV) mainly depend on three factors: *i*) numbers of DSPs devoted to arithmetic operations and their internal configuration; *ii*) the number of accesses to the global memory; and *iii*) the maximum clock frequency of the synthesised design. After synthesis, we were able to run our design at 170.41 MHz. With such clock speed, the 1,332 DSPs provide a raw performance of 454 GFLOPS (DSPs configured for the 32 bit IEEE-754 arithmetic) with an overall power consumption of 30 W (the value also consider power consumption of other components on the board, such as the global memory –up to 8GiB). Great improvement can be obtained by moving on reduced precision arithmetic. The use of fixed point arithmetic allowed DSPs to be split, thus doubling the performance ($\sim$ 1 TFLOPS) without negatively impacting on the power consumption. Further improvements have been obtained by optimising the CNN architecture (YOLO version 2 to this purpose) and training it against the six useful classes for the application (i.e., bicycles, motorbikes, car, trucks, bus, pedestrians). With the described set up (using standard 32 bit IEEE-754 arithmetic), the proposed solution was able to accelerate image detection by a factor of 33$\times$ when compared to pure software execution on CPU (Intel Xeon). The proposed design achieves top-class efficiency levels being able to process up to 50 frames/s (configuring DSPs to use standard 32 bit IEEE-754 arithmetic), corresponding to 1.6 frames/s/W. By configuring the internal DSPs to use fixed point arithmetic, the efficiency almost double, thus also the performance proportionally grow. For comparison, an high-end GPU (Nvidia Titan X) offers an efficiency equals to 0.66 frames/s/W, that is roughly 41% of that sported by our FPGA implementation. This demonstrates the advantages of the proposed architecture.

VII. Related Works

In this section we discuss some works that have been of interest in devising and designing the proposed platform.

Architectural heterogeneity has become common in recent years thanks to the energy saving benefits of using more specialised systems. Historically, heterogeneous solutions have found space in the HPC domain, although accelerators are now becoming common in Cloud infrastructures [12] and CPS. These latter are often deployed in urban contexts as Cloud-connected smart sensors, where they are demanded for computationally intensive tasks to avoid large data transfer to the Cloud back-end [13], [14].

Among the others, computer vision applications became extremely popular since they found opportunity for being accelerated both on edge devices and at scale using discrete accelerators on data centers. To this end, many works targeted efficient implementation of CNNs on GPGPUs, FPGAs [15], and ASICs [16]. Regarding FPGA implementations, many designs rely on creating a large matrix multiply fabric, which is less effective compared to GPGPUs and ASIC implementations. For this reason, we opted for a HLS pipeline which preserve more the CNN algorithm structure and better exploits the underlying data driven resources of the FPGA.

Beside computing elements, an effective chip-to-Cloud solution must optimise also communication links. Moving from power hungry wireless subsystems to more flexible and less consuming ones, requires optimisation of the radio communication subsystem. Low Power Wide Area Network (LPWAN), compared to short-range connections based on WiFi, Zigbee, Bluetooth and cellular connections covers tens of kilometres at the expense of very low data rates [17], [18]. Due to this latter, standard LPWAN are not suitable for video surveillance applications, where video streaming capability is mandatory. Thus, state-of-the-art solutions still rely on expensive mobile networks (e.g., GSM/3G) [19]. Conversely, the WiFi 802.11-based standards (e.g., 802.11a/b/g/n) provide high data rates, operating in unlicensed bands, while covering short/medium distances (up to 500 m).

TABLE II
DATA TRANSMISSION POWER ESTIMATION (MOST CRITICAL SCENARIO FOR EACH TRAFFIC MONITORING USE CASE).

Type of detection	**Statistical number of detection**	**Video duration [*s*]**	**Power consumption P_{case} [*Wh/day*]**
Congestion	50	20	0.82
Wrong-way vehicle	1	60	0.089
Object group counting	10000	0	1.02

There exist several techniques to reduce the consumption of the radio module [20]: *i*) minimising transmission power levels [21]; *ii*) optimising the transmissions scheduling [22]. In this context, reconfigurable antennas play a crucial role, since they provide energy efficiency either alone, or in synergy with other energy saving approaches [23]. Higher antenna gain can be obtained by controlling the steering of the beam [24] through a configuration process. Our proposed solution aims at further improving energy efficiency of a traffic monitoring platform through the implementation of a flexible and optimised directive (i.e., reconfigurable) antenna thanks to a complete software-defined radio (SDR) approach.

VIII. CONCLUSIONS

The proposed chip-to-Cloud architecture ambitiously integrates different technologies into a globally energy efficient video surveillance platform optimised for traffic monitoring applications. Heterogeneous and (ultra-)low power technologies have been exploited to achieve high levels of energy efficiency, along with the capability of the system to run complex image processing tasks. Further, the proposed architecture allows to efficiently transfer data from edge devices to data center, by exploiting an innovative reconfigurable wireless subsystem. The implementation of state-of-the-art CNNs both on Cloud-edge and Cloud server for image analysis in the context of traffic monitoring, as well as its validation in field, demonstrated the great benefit of the proposed system in real contexts.

ACKNOWLEDGMENTS

This work is supported by the OPERA project, which has received funding from the European Union's Horizon 2020 Research and Innovation programme under the grant agreement No. 688386. The authors also thanks all the OPERA partners for their useful comments and feedbacks.

REFERENCES

[1] *OPERA EU project*, accessed October 10, 2018. [Online]. Available: http://www.operaproject.eu/

[2] R. Montella, C. Ferraro, S. Kosta, V. Pelliccia, and G. Giunta, "Enabling android-based devices to high-end gpgpus," in *International Conference on Algorithms and Architectures for Parallel Processing*. Springer, 2016, pp. 118–125.

[3] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.

[4] G. Giordanengo and et al., "Energy efficient system for environment observation," in *Proc. of the Int. Conf. on Complex, Intelligent, and Software Intensive Systems (CISIS), Torino, Italy*, July 2017, pp. 987 – 999, doi:10.1007/978-3-319-61566-0-93.

[5] S. Ciccia, G. Giordanengo, and G. Vecchi, "Open-source implementation of an ad-hoc ieee802.11a/g/p software-defined radio on low-power and low-cost general purpose processors," *Radioengineering*, vol. 26, no. 4, pp. 1083 – 1095, 2017, doi:10.13164/re.2017.1083.

[6] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv*, 2018.

[7] E. Commission, "Photovoltaic geographical information system (pvgis)." [Online]. Available: http://re.jrc.ec.europa.eu/pvgis/

[8] ——, "Calculation of pv power output." [Online]. Available: http://re.jrc.ec.europa.eu

[9] E. Solar, "Et m53620." [Online]. Available: https://www.bajadanewenergy.com

[10] Sonnenschein, "Battery model a512/10 s*." [Online]. Available: http://www.eurosep.com

[11] G. Desoli and et al., "14.1 a 2.9 tops/w deep convolutional neural network soc in fd-soi 28nm for intelligent embedded systems," in *Solid-State Circuits Conference (ISSCC), 2017 IEEE International*. IEEE, 2017, pp. 238–239.

[12] S. P. Crago and J. P. Walters, "Heterogeneous cloud computing: The way forward," *Computer*, vol. 48, no. 1, pp. 59–61, 2015.

[13] A. Traber and et al., "Pulpino: A small single-core risc-v soc," in *3rd RISCV Workshop*, 2016.

[14] N. Ickes and et al., "A 10 pj/cycle ultra-low-voltage 32-bit microprocessor system-on-chip," in *ESSCIRC (ESSCIRC), 2011 Proceedings of the*. IEEE, 2011, pp. 159–162.

[15] J. Qiu and et al., "Going deeper with embedded fpga platform for convolutional neural network," in *Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 2016, pp. 26–35.

[16] R. Andri, L. Cavigelli, D. Rossi, and L. Benini, "Yodann: An architecture for ultralow power binary-weight cnn acceleration," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 1, pp. 48–60, 2018.

[17] F. Adelantado, X. Vilajosana, P. Tuset-Peiro, B. Martinez, J. Melia-Segui, and T. Watteyne, "Understanding the limits of lorawan," *IEEE Communications Magazine*, vol. 55, p. 34 40, 2017, doi:10.1109/MCOM.2017.1600613.

[18] H. Wang and A. O. Fapojuwo, "A consumer transceiver for long-range iot communications in emergency environments," *IEEE Communications Surveys & Tutorials*, vol. 19, p. 2621, 2017.

[19] Y.-W. Kuo, C.-L. Li, J.-H. Jhang, and S. Lin, "Design of a wireless sensor network based iot platform for wide area and heterogeneous applications," *IEEE Sensors Journal*, vol. 18, pp. 5187 – 5197, 2018, doi:10.1109/JSEN.2018.2832664.

[20] Z. Zhou and et al., "Energy-efficient optimization for concurrent compositions of wsn services," *IEEE Access*, vol. PP, pp. 657 – 660, Sept. 2017, doi:10.1109/ACCESS.2017.2752756.

[21] D. Basu and et al., "Energy efficiency comparison of a state based adaptive transmission protocol with fixed power transmission for mobile wireless sensors," *Journal of Telecommunications System and Management*, vol. 6, 2017, doi:10.4172/2167-0919.1000149.

[22] P. Le-Huy and S. Roy, "Low-power 2.4 ghz wake-up radio for wireless sensor networks," in *Proc. of the IEEE Int. Conf. on Wireless and Mobile Computing, Networking and Communications, Avignon, France*, Oct. 2008, pp. 13 – 18, doi:10.1109/WiMob.2008.54.

[23] G. Manes and et al., "Energy efficient mac protocols for wireless sensor networks endowed with directive antennas: a cross-layer solution," in *Proc. of the IEEE Radio and Wireless Symposium, Orlando, FL, USA*, Mar. 2008, pp. 239 – 242, doi:10.1109/RWS.2008.4463473.

[24] S. Ciccia, G. Giordanengo, and G. Vecchi, "Reconfigurable antennas for ultra low-power radio platforms based on system-on-chip," in *12th EUropean Conference on Antennas and Propagation (EUCAP), London, UK*, April 2018.